CHAPTER

14

BACKING UP SYSTEM DATA

After reading this chapter and completing the exercises, you will be able to:

- Understand the issues surrounding backups and back-up strategies
- Discuss hardware and software issues related to backing up Linux data, such as back-up devices and storage media choices
- Use popular back-up utilities such as `tar`, `cpio`, and graphical back-up utilities

In the previous chapter you learned about managing the printing services on Linux. You learned how to create a printer configuration file and how to manage print queues using command-line or graphical utilities. You also learned how to print to remote print servers and to printers attached to computers running other operating systems.

In this chapter you will learn about backing up data stored on a Linux system. You will learn how to plan effective back-up strategies for different types of systems and environments. You will also learn about the hardware components, back-up media, and software utilities available to help you create and manage backups of your data.

BACK-UP STRATEGIES

As you learned in Chapter 9, no matter how many precautions you take, all computers are subject to failure. Thus, making backups of essential files is a form of insurance. In its simplest form, a **backup** is nothing but a copy of data on a computer system. However, backing up thousands of files owned by dozens or hundreds of users is not a simple process. But because the time and equipment needed to make backups are minimal compared with the costs associated with lost data, nearly all organizations regularly back up the files on their computer systems according to an established back-up plan. A **back-up plan** is a written document that outlines when, how, and, perhaps, why various files and file systems will be backed up, stored, and—when necessary—restored. As you might guess, implementing the back-up plan normally falls to the system administrator.

Among other things, the back-up plan should specify the type of back-up media to be used. The term **back-up media** refers to the device that stores the backed-up data, such as a tape cartridge (the most common format), writeable CD, or even a floppy disk. The back-up plan should also specify how lost data should be restored. The term **restore** refers to the process of copying data from a back-up location (for example, a tape cartridge) onto the file system where that data is normally used, and from which it was unintentionally lost.

Because of the complexities involved, developing a back-up strategy that works well in any organization is an ongoing process. As a system administrator, you can expect to work with numerous existing and new computer systems, a variety of applications and data storage needs, and computer users whose preferences and actions are rarely predictable. The following sections address some of the questions that you should consider when formulating a back-up plan.

Asking Initial Questions

Some of the initial questions that you'll want to consider as you formulate a back-up plan include the following:

- *What files should be backed up?* You might initially think that everything on the system needs to be backed up. Although that's an admirable goal, time and cost restrictions might make it impractical. You can evaluate various parts of your system to determine what data is easily restored from CD, such as the operating system or an application. If you are short on resources, these items can be re-created (and then reconfigured) from their original sources rather than from a backup that you create.
- *Who will back up files?* As mentioned previously, this normally falls to the system administrator. You may, however, decide that users on a networked system have some responsibility. Perhaps users should be informed that only data placed in a certain directory area will be backed up each night. Or a system administrator may share the responsibility for backups with a colleague, either to reduce the work burden on one person or to make backups more accessible in case they are needed for restoring data.
- *Where are files located?* You probably know offhand where most of the different types of data are located on your Linux system. A more thorough approach can help you see which specific directories on the system are being actively used, which contain data that is easily reconstructed, and which hold temporary files that don't warrant the effort of a regular backup. These are just three examples of the categories you might assign to parts of your system as you review the various file systems and devices that store data.
- *How should backups be performed?* The answer to this question may be determined by the equipment you purchase, as well as by how your organization operates its computer systems. Many system administrators must back up data during non-work hours. This process can be automated in most cases using a cron job (see Chapter 12). You might also want certain events to trigger a regular backup, or a different type of backup than would normally occur. For example, you might want to back up the entire system before installing new hardware devices such as SCSI adapters.

- *Must you be able to restore data within a specific period of time?* When a problem occurs (and it will), several factors affect how rapidly you can restore lost data to the system. These factors include the size and location of the lost files and the media format on which the back-up data was stored. Your backup plan should reflect the importance of timing in your organization. In some organizations, the ability to restore lost data immediately is essential. In others, speed may not be quite as critical.

A well-designed back-up plan will make it easy and convenient for you to regularly back up system data and restore files. Ideally, your back-up plan should prevent the headaches associated with having to locate files and figure out how to reconstruct damaged or lost data.

Determining the Value of Data

As with creating redundant systems, your back-up strategy should be based at least partially on the value of the data that you are backing up. The more expensive data is to create, acquire, or refine, the more you should spend to protect its integrity. Some data may only be valuable to one person in an organization, but if that person's time is required to re-create any data that is lost, the data still has value to the entire organization.

As an example, a study of the value of data held by an organization might determine that a given set of files required 4,000 hours of work by the employees of the firm to create. A different estimate might state that the data could be re-created given current experience and facts in about 2,000 hours. If the average wage of the employees involved in the project is approximately $40 per hour, the data would have a value of $80,000. But the study doesn't end there.

The estimate of 2,000 hours—about one work-year—is based on an experienced employee re-creating the data. If that well-trained employee spends time re-creating lost data, what current work will he or she not be able to do? This is called the opportunity cost. The employee might forgo a project worth many times $80,000 in order to re-create the lost data. Opportunity cost extends even further. How was the data that was lost going to be used? Was it part of a multimillion dollar advertising campaign? Or perhaps a financial merger? A great deal of money may be lost because the data is unavailable when needed. Even if $80,000 can be invested to re-create it, the moment of opportunity when the data was needed may be past.

TIP

This discussion doesn't address the anger or low morale of an employee who must re-create a project that was partially or completely finished. These are also key factors in any organization.

The following list summarizes questions to ask when determining the value of data:

- How many hours of effort were spent creating the data?
- How many hours of effort would be required to re-create the data?
- How much inherent value does the data contain for the operation of the organization?

- Is the data irreplaceable?
- Is the data time critical to a current project?

These considerations are similar to those raised in the discussion of hardware redundancy and fault tolerance in Chapter 9. The decisions you make as a system administrator are also similar to those you might make when evaluating your system's hardware: if data is worth millions of dollars to your organization, don't hesitate to spend $50,000 to $100,000 to protect that data. By answering the questions in the preceding list, you may be able to convince company officers or supervisors that the expense is warranted. With the right hardware and software tools, you'll be well prepared to secure the information that you safeguard as a system administrator.

Determining When to Back Up Data

Once you have created an initial backup or archive of important data, the question of how often to refresh the backup arises. Having at least one backup of data is better than having none at all, but data changes frequently in most organizations. Continually backing up the latest information stored on the system is a critical part of most system administrators' jobs.

The question of when to back up data is related to how valuable the data is to an organization. You need to start by asking, "How often does the data change?" Another good question to ask is this: "Do changes to the data affect the value of the data?"

The answers to these questions vary, depending on which part of your Linux system you are evaluating. The operating system itself probably changes very little after your initial configuration efforts. Applications installed on the system are also unlikely to change regularly. By contrast, user data, log files, and other items change rapidly and are normally the focus of back-up efforts. This data constitutes the daily work of others within your organization. By maintaining regular backups, no one is ever likely to lose more than a few hours worth of work, even if the entire system crashes or a hard disk is destroyed.

Several back-up strategies are commonly used. You can select a strategy based on how often data on your system changes and how valuable or critical each incremental piece of data is. The following discussion describes a widely used back-up strategy for Linux.

A Linux Back-up Strategy

Various strategies have evolved among Linux users to balance the need for a complete backup of data at all times with the need for convenience in creating and maintaining backups. The method described here is accepted as standard for most Linux and UNIX systems. You can adjust the time frame according to how often the data on your system changes.

Using Back-up Levels

To understand this back-up method, you need to understand the concept of a back-up level. A **back-up level** defines how much data is to be backed up in comparison with another back-up level. A back-up operation at a given back-up level copies all of the data that has changed since the last backup of the previous level. For example, a backup at level 1 stores all files that have changed since the last level 0 backup; a backup at level 2 stores all files that have changed since the last level 1 backup. A standard system might operate with three levels, as described here:

- Level 0 is a full backup. Everything on the system is backed up. Suppose for this example that a level 0 backup is performed on the first of every month.
- A level 1 backup is done once per week. Every file that has been modified since the last level 0 backup (on the first of the month) is included in the level 1 backup. This is referred to as an incremental backup.
- A level 2 backup is done each day. Every file that has been modified since the first of the week (the last level 1 backup) is included in the level 2 backup. Like a level 1 backup, this is considered an incremental backup.

Figure 14-1 illustrates the three-level backup just described.

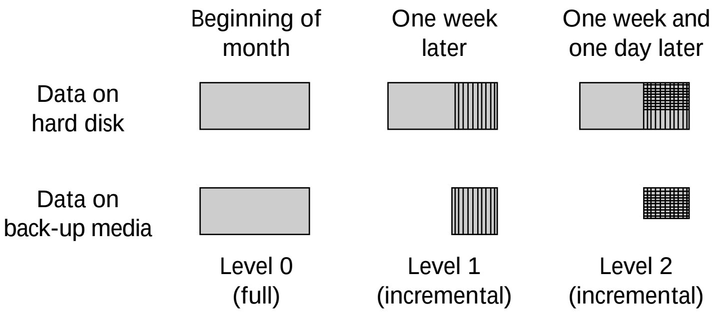

Figure 14-1 Back-up levels

The times associated with back-up levels are arbitrary, though a level 0 backup is normally a full backup in which every file is backed up. Beyond that, you can assign any time frame that you choose to each level; the point is simply that each level records all the changes since a backup of a previous level.

The advantage to using back-up levels is that you can back up data frequently—meaning very little work is lost if a system fails—but you don't have to back up the *entire system* each time you do a backup.

14

Restoring a File from a Three-Level Backup

Now consider how a system administrator would restore a file that a user had deleted and needed help recovering. The user can't recall when the file was last modified, but it was "recently." The system administrator follows these steps to locate the file:

1. Check the most recent level 2 backup. If the file is there, it was changed in the last day. This backup probably doesn't include very many files compared to the size of the entire system, so it's easy to search for a file. If the file isn't there, then it wasn't modified in the last 24 hours, so proceed to Step 2.
2. Check the most recent level 1 backup. If the file is there, it was changed sometime after the first of the week, but not in the last 24 hours. This backup contains more files, so it takes a little longer to search. If the file is not found, proceed to Step 3.
3. Check the most recent level 0 backup. The file will always be located on this backup because a full backup includes every file on the system. But searching through this backup may be time consuming because it is fairly large.

TIP

Back-up media such as tape drives and optical disks always have directories of their contents to help you locate files as rapidly as possible, but a tape cartridge must be rewound to the place where the file is stored. As a result, restoring a single file from a tape cartridge can still be time consuming.

You may wonder why you shouldn't start searching for the file in the level 0 backup. You should always start with the most recent backup in order to find the most recent version of a file. If the file had been altered since the first of the month, the most recent copy of the file will not be on the level 0 backup. Hence you should start with the most recent backup (level 2) to see if the file is located there.

Advantages to the three-level back-up method include:

- Creating the level 2 daily backups requires little of the system administrator's time because few files are altered on any given day.
- No user will ever lose more than a single day's work because the changes in the file system from each day are recorded in a level 2 backup.
- Files that rarely change are still backed up and available, but don't require daily maintenance by the system administrator.

Some back-up utilities explicitly use the term back-up levels to refer to how data is backed up and how back-up media are tracked. The concept can be applied to any utility, however. For the system to work well, you need to keep careful records and label back-up media clearly.

In the event that an entire system must be restored using a set of back-up media that have been prepared using the three-level method, a system administrator would follow this procedure:

1. Restore everything from the latest level 0 backup.
2. Restore everything from the latest level 1 backup.
3. Restore everything from the latest level 2 backup.

Figure 14-2 illustrates how this procedure will result in all of the latest information being included in the restored file system. (Compare the back-up levels pictured in Figure 14-1 to the restore operation pictured in Figure 14-2.)

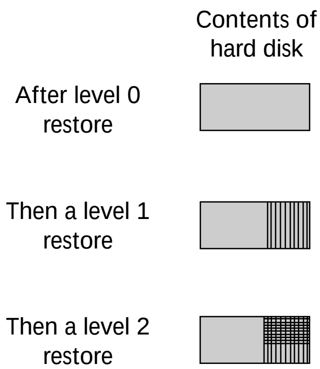

Figure 14-2 Restoring data from a set of back-up media with levels 0, 1, and 2

Managing and Storing Back-up Media

As you create a back-up plan that specifies back-up levels and times appropriate to your needs, you must determine how many back-up media you will need (disks, tapes, cartridges) for each level. That is, a level 0 full backup may require five tape cartridges, but a typical level 2 backup requires only a single cartridge (because relatively few files are modified each day). As an example, the three-level backup described previously might include the following:

- Three months of level 0 backups; each requiring 5 tape cartridges, for a total of 15.
- Five weeks of level 1 backups (some months have five weeks); each requiring 3 tape cartridges, for a total of 15.
- Five days of level 2 backups (you might need seven days if your organization runs seven days per week); each requiring 1 tape cartridge, for a total of 5.

You would therefore need a total of 35 tape cartridges. Figure 14-3 illustrates this arrangement. The importance of carefully labeling each tape cartridge cannot be overstated. If you can't identify which back-up media is the most recent of any given level, much of your back-up efforts will be useless when a serious problem arises.

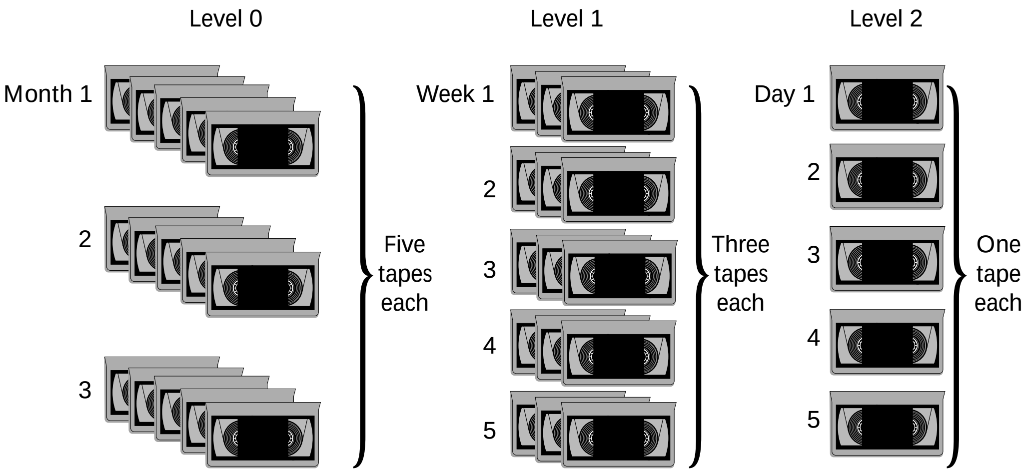

Figure 14-3 Multiple tapes used for a three-level back-up plan

Most organizations would store one set of the monthly (level 0) back-up media and perhaps the most recent weekly (level 1) back-up media off-site. The strategy for off-site storage depends on how critical data is and how often someone wants to take the responsibility of carrying the back-up media to the chosen secure location (such as a bank vault).

> **TIP** Most back-up media are designed to be used repeatedly, like a floppy disk. For example, a rewriteable CD can be used about 1000 times, according to the manufacturer. But you should nevertheless plan on a scheduled life for back-up media, so that you avoid problems with deteriorating, outdated products.

Using the plan just outlined, you could reuse the same set of level 1 weekly tape cartridges each month, starting with the oldest one. The same applies to the level 2 tape cartridges. For example, on any Wednesday afternoon, you should have five level 2 tape cartridges containing the following:

- Last Wednesday's backup, which you will overwrite this evening with new data
- Last Thursday's backup, which you will overwrite tomorrow evening with new data
- Last Friday's backup
- Monday's backup (from two days ago)
- Tuesday's backup (from last night)

In addition to being fairly easy to manage, this system provides data redundancy. If you have a problem and need to restore a file on this particular Wednesday, you first check the Tuesday backup that you made last night. If a problem occurs with that tape cartridge, you can also check Monday morning's level 1 backup, Monday evening's level 2 backup, or even last Friday's backup. A user may lose more work if you cannot use the most recent backup, but the user is unlikely to lose everything, because many copies of the file exist, created at different times.

Backing Up the Root File System

The root file system requires special attention in your back-up plan because it contains the tools that you normally use to restore damaged data, such as a deleted file or data from a corrupted hard disk partition. You must therefore think about how to respond if the root file system is damaged, either by a hard disk failure or by corrupted configuration files that prevent you from booting the Linux operating system kernel.

Chapter 9 described how to create a rescue floppy or a boot floppy. That disk, which you can use to boot the system in an emergency, should contain the files necessary to restore the contents of the root file system from your back-up device. These files might include:

- The kernel modules needed to access the back-up device (such as SCSI modules)
- Other kernel modules needed to access the device where the root file system is stored
- Configuration files needed to set up access to the back-up device
- Information such as file indexes that are needed to locate the correct data and restore it from back-up media

As you prepare a back-up plan, you'll want to consider the hardware and software that you'll use to implement that plan. The next part of the chapter describes some key issues you face in making hardware and software choices.

HARDWARE AND SOFTWARE ISSUES

Once you have determined why, when, and how you want to back up your Linux system, you must determine the best tools to use to get the job done. Linux includes all the necessary software utilities for many back-up tasks. You can also purchase commercial back-up software. Both of these options are described later in this chapter.

Many different hardware devices are available for backing up data. The next few sections provide a review of the different options available to you.

Choosing Back-up Media

The size of hard disks in standard PCs is growing very rapidly. Whereas a 500 MB hard disk was considered huge just a couple of years ago, hard disks with 50 GB—100 times that amount—are now available for well under $1000. Storage space is often measured according

to its cost per megabyte. For example, if a 16 GB hard disk costs $400, the cost per megabyte is about 2.5 cents. Similarly, if a tape cartridge used for backing up a system costs $79 and holds 20 GB, the cost per megabyte is 0.38 cents per megabyte.

When you back up your data, you will normally have multiple copies of the data that was backed up at different times. Back-up media such as tape cartridges generally cost much less than a hard disk or other similar device, but you must purchase multiple tape cartridges to back up the system.

Unfortunately, back-up devices have not kept pace with the growth in capacity of hard disks, though many different formats and devices are available for system backups, as described here. The following paragraphs review the different back-up media (and corresponding devices) that you might consider for backing up your Linux system.

Magnetic Media

Several types of magnetic media are occasionally used for specialized back-up needs.

- *Floppy disks:* although you might be surprised to see this item listed, floppy disks are a great way to back up small, sensitive pieces of information. For example, a boot disk, a rescue disk, a firewall or other server configuration, and other similar data can easily be copied to a floppy disk. The disk is inexpensive, easily transported, and easily stored. Just be certain to label the disk and move the write-protection tab over so you don't erase the floppy disk. An important disadvantage of floppy disks is that they are fragile. You should maintain multiple floppy disk copies of any critical data and check the integrity of the disks regularly.
- *RAID hard disks:* most of the data that you want to back up is already on a hard disk. It doesn't make sense to rely on long-term data storage located on another hard disk—even a RAID array—if the same vulnerabilities apply to that device as to your main hard disk. On the other hand, storing a back-up copy of crucial data from several locations on a centralized RAID array is a useful way to maintain an online backup—that is, a backup of the data that is still available if one of the hard disks becomes unavailable. In general, however, don't plan your back-up strategy around this sort of thing. Instead, look to removable devices such as tapes and removable cartridges.
- *Removable media:* many types of specialized cartridge storage devices are now available. These include Syquest cartridges, Floptical and similar devices that store a large amount of data on a small disk similar in size to a floppy disk, and various products such as the Zip and Jaz cartridges from Iomega. The data capacity of these cartridges continues to rise. The latest Jaz cartridges hold 2 GB each.

Removable media, the last item in the preceding list, have several advantages, including the following:

- Random, immediate access to any point on the media, similar to a hard disk
- The ability to expand storage by purchasing additional cartridges

- Relatively easy access to the back-up device—most are treated like a standard hard disk, formatted with the `ext2` file system, and mounted normally.

Removable media also have disadvantages, such as:

- High cost per megabyte of storage
- Proprietary formats (compared to most tape backups), which may mean difficulty obtaining new cartridges in the future and lack of support from other vendors

Optical Media

Optical media used by devices such as writeable CD drives and DVD drives are an attractive back-up choice. Advantages of optical media include:

- Their large storage capacity is sufficient for many needs.
- Storage media are very low cost.
- Storage media are widely available.
- Optical media are easily exchanged with vendors, customers, or other organizations.

Standard CDs, in particular, are a valuable method of exchanging large amounts of data with suppliers and also of easily creating data archives. Because a single writeable CD costs less than a dollar, it is cost effective to back up key data files regularly on a CD and to have a set of back-up CDs stored with snapshots at various times. Rewriteable CDs, which you can update in the same way that you update data on a hard disk, cost a little more but provide more flexibility. The capacity of a CD is only about 640 MB. That's not much compared with the data stored on an entire hard disk, but it's often sufficient for backing up an entire project directory, graphics archive, programming project, or operating system.

DVD drives, which are increasingly popular for watching movies on a computer, also come in a writeable format called DVD-RAM. A DVD-RAM cartridge (costing under $50) holds about 5.2 GB of data. Drives are inexpensive as well. For data sets too large for writeable CDs, the low cost and wide use of DVD make it an attractive choice.

Tape Cartridges

Tape drives are the workhorses of most computer back-up efforts. Tape drives are fairly inexpensive, as is the media (tape cartridges). Many formats are available, but in general, data capacities have kept pace with that of hard disks. Thus you can purchase a tape drive that will record 8, 40, or even 100 GB on a single tape cartridge. All such cartridges are priced under $100, with the smaller capacities costing far less. If you need to back up large amounts of data, such as hundreds or thousands of gigabytes, you should consider special tape cartridge jukeboxes or high-end digital tape formats available from major device manufacturers such as IBM and Hewlett-Packard. The term **jukebox** refers to a back-up device that holds multiple back-up media (such as multiple tape cartridges or writeable CDs) and can switch between them without assistance from a system administrator.

Tape drives are available in a variety of formats, and new formats seem to appear each year as manufacturers rush to keep up with growing capacities and speedier computers. Manufacturers of the latest tape drives claim storage capacities of up to 200 GB on a single tape cartridge; others claim data transfer rates in excess of 200 MB per minute. Explaining the features of a diversity of tape formats is beyond the scope of this book, but the information that follows provides enough basic details to familiarize you with the formats you're likely to see.

Keep in mind that tape cartridges can accommodate different methods for storing data, depending on the tape drive you use. This is similar to a regular 3.5-inch floppy disk, which can be formatted with either an MS-DOS, Macintosh, or Linux file system.

When reviewing the great number of tape devices on the market, you may feel overwhelmed by the alphabet soup of formats, companies, and product names. The following list describes some major tape cartridge device types and data formats.

- Digital Linear Tape (DLT) is a half-inch-wide tape inside a cartridge. The tapes store up to about 40 GB and are considered highly reliable. Quantum is considered the leader in DLT technology, but many others, such as StorageTek, also use DLT.
- Linear Tape-Open (LTO) is an open tape standard used by Hewlett-Packard, IBM, and Seagate (a prominent hard disk manufacturer). Many companies are currently planning devices based on this high-capacity format.
- Helical-scan tape drives write data onto a thin tape—either 4mm or 8mm. This storage format is the same method used by videotapes for recording movies. Figure 14-4 illustrates how a helical-scan device stores information by writing short, angled strips of data on the tape. Helical-scan tapes (usually the 8mm size) are used in several newer tape formats as described below.
- Advanced Intelligent Tape (AIT) is a format developed by Sony. Each AIT cartridge contains a memory chip that is used to increase the efficiency of data access. Sony plans to release a revised version of AIT every two years, with a doubled storage capacity and data transfer rate in each new version. AIT-3 tapes are expected to hold 100 GB and transfer data at about 720 MB per minute. (AIT-3 devices are not available at the time of writing.)
- VXA is a technology developed by the Ecrix company. The VXA format attempts to overcome some of the technical limitations that most other standard formats face. For example, VXA avoids stopping and starting the tape drive while waiting for the computer to send more data by using a variable speed tape drive and organizing data into packets rather than a single stream, as most formats use.
- Travan tape drives are widely used and are manufactured by many different companies. They do not have high capacities—10 to 20 GB is standard—but they have a longer history of reliability than many of the newer formats. Travan uses the QIC tape cartridge format.

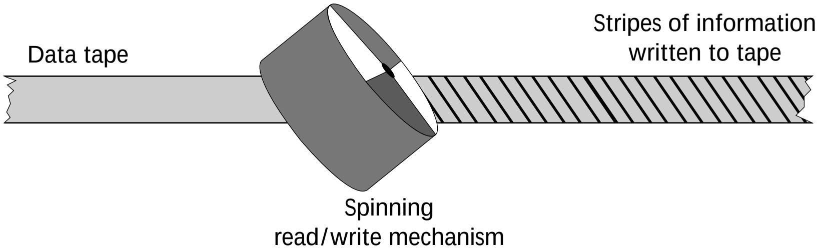

Figure 14-4 Using the helical-scan method to write data on a 4mm or 8mm tape

After reviewing the available formats, media, and devices in this section, you can use the information in the next section to help you determine what factors to consider when choosing a back-up device.

Comparing Devices

Deciding among all of the available back-up devices and technologies is challenging. System administrators who are creating a new system from scratch and need to store large amounts of data may be able to focus on the latest technology for high-capacity, high-speed tape drives. Other system administrators may be more concerned with sharing copies of data between several existing computers using a CD drive, and so may opt for a low-cost CD recorder. Still other administrators may be forced to purchase new devices that use older technologies simply to keep costs low or maintain compatibility with existing systems, even though this means much more work to maintain numerous back-up media. These are just three examples of the varying requirements that have led manufacturers to develop so many different devices and formats.

In most cases, the cost of the back-up device and the back-up media is an issue, at least peripherally. Although you should keep in mind the discussion at the beginning of this chapter regarding the value of an organization's data, managers who control budgets will still expect you to be as frugal and wise as possible with an organization's money. The cost of the various back-up device options is affected by several factors, including the following:

- *What interface is used to connect the device to the computer?* The interface is usually SCSI (fast and expensive), IDE (common and inexpensive, but slower), or parallel port (great for special applications and inexpensive, but quite slow compared to IDE and SCSI).
- *How recent is the format?* The more recently developed formats are more expensive. They generally hold more data, however.
- *How much data can one disk or cartridge hold?* The same media may be able to hold slightly different amounts of data when used in different devices. The difference in storage is unlikely to be more than 20%, however.

- *Is the device from a name-brand manufacturer?* As with everything else in the computer industry, buying a product from a company such as IBM or Hewlett-Packard generally costs more than buying from a start-up or relatively unknown company. The start-up company may support newer technologies, however, and may actually have better products. New companies often must compete on price until their quality or technology is recognized.
- *Does the device have special features?* The most common of these is an automounting or jukebox feature, which allows you to load a number of disks or cartridges so that the device can create a multivolume archive without user intervention. These devices are always much more expensive than a device supporting a single manually inserted disk or cartridge. They are also more subject to breakdown because of the additional mechanisms required to automate media handling.

Among all the devices available, your decision may be guided by many things. The following list presents a summary of factors that you should consider in selecting a device to fulfill your back-up strategy. This list is not exhaustive, but it should give you a good start at considering which device will be best for your needs, and also understanding why so many different devices are available in the market.

- *How much does the hardware device cost to acquire?* A quick survey on a major hardware supplier site such as *www.warehouse.com* will show you prices ranging from $150 for a used 2 GB Iomega Jaz drive to more than $10,000 for a high-capacity, name-brand jukebox tape cartridge system.
- *How rapidly does data transfer from the computer system to the back-up media?* This is less important if you intend to back up using a scheduled cron job in the middle of the night. It may suddenly become important again if you need to restore a large amount of data with many people waiting while you do it. Data transfer rates are usually measured in megabytes per minute (MB/min). For example, a 14 GB tape drive might advertise a data transfer rate of 78 MB/min, meaning that an entire 14 GB tape could be filled in about three hours. Faster transfer rates generally cost more, but the data transfer rate that you actually achieve is based on many factors, such as the speed of your CPU, the quality of your system board components, and the type of interface used to connect to the back-up device.
- *Is data randomly accessible?* In other words, is it easy to back up a single file or set of files without restoring or going through an entire archive set?
- *Can the device you choose perform very large backups using an autoloader or jukebox mechanism?* If it can't, you may always need to be present during system backups to switch media when one is full.
- *How much does media for the device cost?* You can expect media costs to be greater in the long term than the hardware cost if you use a device for several years. Determine the media costs based on your planned back-up strategy, with replacement media purchased regularly as recommended by the manufacturer to avoid storage errors.

- *Does the organization already own some back-up hardware?* Is the cost savings for not purchasing new hardware worth using the existing hardware if it relies on older or obsolete technologies? When dealing with this issue, you will often hear the term legacy systems. **Legacy systems** are systems that an organization already owns. Working with existing systems is a common concern when organizations plan new computer hardware or software acquisitions.
- *How recent is the technology of the device?* Some older devices are still very viable and stable, but may be difficult to locate media for. If you have a nine-track tape reel in your office, you may have to work with a special service bureau to read the tape because these devices are rarely used now. On the other hand, very new technologies may not have proven themselves cost effective or technologically sound. For example, some tape formats had problems when first released to the public because magnetic material flaked off of the storage tapes.
- *Does the device rely on an industry standard that many vendors support, or is it specific to one vendor?* If the device is only supplied by one vendor, can you rely on that vendor to be around for a while, or could that vendor change formats or discontinue a product, leaving you with outdated products or the prospect of retooling your back-up plans?
- *How long is the media life?* This may not be a big issue if you are working with daily backups, but most organizations maintain some sort of archival backup of company financial records, personnel records, computer program source code, and other electronic assets. The media that these assets are stored on should last long enough so that you are not required to make an updated copy of massive amounts of data every two years. Table 14-1 shows the anticipated life of some key materials. (Paper and microfilm are included in the table for comparison, not because you would use them as back-up media.) Note that the industry's experience with these technologies does not extend past their supposed useful life. We *know* that paper can last 500 years. No one really knows if CDs can last 30 years, because they haven't existed that long.

Table 14-1 Comparison of Media Life

Media	Approximate useful life (before data loss potentially occurs)
Archive-quality acid-free paper	500 years
Microfilm	100 years
CD-ROM and similar optical media	5–30 years, depending on media quality
Hard disks and similar magnetic media	10–20 years, depending on media quality
Reel-to-reel data tapes	15–25 years, depending on media quality
Tape cartridges (QIC, 4mm, 8mm, etc.)	5–10 years, depending on media quality
3.5 inch disks	2–5 years

- *How robust is the media?* Can they be dropped? Can they handle the environmental conditions that are part of your working area (heat, humidity, dust)? Most media formats are quite robust, but if you work in a factory or outdoor environment, you should consider these factors.
- *Is the media easily transportable (if this is a requirement of your organization)?* Most system administrators keep the majority of back-up media near the systems that contain the original data. This makes it convenient to restore data if a problem occurs. But it's also a good idea to take at least one copy to another location in case a fire or other problem destroys the back-up media located nearby. Many organizations have a strict policy about regularly taking a data backup to a bank vault or other secure off-site location.
- *Do you need to exchange data with other organizations, or will you rely on service bureaus to help you process or recover data from archive media?* In this case you should consult several service bureaus and select a media format and device that are widely available.
- *How reliable is the hardware device?* An unreliable or faulty back-up device can corrupt back-up media so that no device can read them. Even if a hardware problem doesn't corrupt media, a breakdown can interrupt your scheduled back-up times or delay restoring data when a problem occurs.

Once you have selected a back-up device and media format, you are almost ready to implement your back-up plan. But a few additional issues still remain to be resolved. These are discussed in the next section.

Verification, Permissions, and Other Issues

As the saying goes, "Trust everyone, but lock your doors." The equivalent tactic when backing up your system is to verify your backups on a regular basis. Verifying a backup is sometimes done as part of a back-up utility, as described later in this chapter, but you can always perform your own verification using steps such as these:

1. Pick a back-up tape or disk, either at random or according to a reasonable plan. For example, you might decide to test a randomly chosen level 1 back-up tape once per week.
2. Check the file listing on the tape by querying for the contents of the back-up media. (This would be equivalent to using the `ls` command to see the contents of the back-up media. With some media you can actually use the `ls` command, with others you'll need to use a back-up utility.)
3. Restore a randomly selected file to the `/tmp` directory of your Linux system, just to be certain that the data in the file can be retrieved and reassembled without errors. If possible, do this step immediately after backing up data (on your regular schedule), and then compare the file you restored with the original file that you backed up to see that the size and contents match.

When you back up data, exactly what information is backed up? Does the backup include the contents of each file? What about the owner and file permissions associated with each file? Many times a system administrator will have problems after restoring a large number of files because the owner and group assigned to files and directories, or the file and directory permissions, are not stored as part of the backup. The consequences of this can range from no one being able to access his or her data once it is restored after a system shutdown, to everyone being able to access everyone else's data on the system, including the system configuration files. You'll have to decide which is worse in your organization.

Back-up utilities normally include options to maintain or ignore file ownership and permissions. Normally you will want to maintain this information and check it carefully when you verify your backups by restoring selected files.

Another issue related to how you choose to use back-up utilities involves the compression feature that most utilities provide. Tape drives typically list a standard capacity and a compressed capacity; back-up commands include options to compress data as it's being archived. Should you use these features? Probably so, but you should also be aware of their limitations. By definition, when you compress data you remove the redundancy from it. That is, compressed data can be re-created in its original form by adding back the redundant information using an established set of rules.

To understand compression better, consider this example. When you see the words "hllo my nm is Nchlas," you can probably understand their meaning even though part of the information is missing. The missing information is redundant—it's not needed for you to understand the sentence. You can also use standard rules (English grammar and spelling) to reconstruct the original sentence: "Hello my name is Nicholas."

The danger with using compression is that with all the redundancy removed from a set of information, all of the information and rules are needed in order to reconstruct the data. For example, if you don't speak English well, English words with missing letters are difficult to decipher. In the same way, if even a small part of some compressed data is lost, the original cannot be easily reconstructed. By leaving the redundancy in the data that you back up, you might make it easier to fix any problems that occur on back-up media.

All modern back-up media formats are highly reliable, but when age, environmental factors like heat and dust, and regular wear and tear are working against the data you have carefully saved, you should consider whether compression is always necessary.

USING BACK-UP UTILITIES

Many utilities are available to back up data from a Linux system in a secure and organized way. The most widely used of these utilities are the old UNIX standbys `tar` and `cpio`. Some of the other utilities use these programs in the background while they present a graphical interface to make configuration and selection of back-up options easier. Popular commercial back-up utilities include features such as tracking tapes for you, keeping online indexes of each backup that you have performed, and automating schedules for unattended backup (similar to the options provided by the `crontab` command).

The following sections outline basic information about using these back-up utilities. Although a complete discussion of `tar`, `cpio`, and commercial tools is not presented here, you should understand enough to use these tools for basic backups and to locate more exhaustive information when needed.

Using `tar` and `cpio`

The name **`tar`** stands for *tape archive*; it is the oldest of the back-up tools for UNIX. The **`cpio`** command (for *copy in and out*) is newer and includes additional features compared to `tar`. `cpio` also reads `tar`-formatted files. Both `tar` and `cpio` can create archive files, such as the `.tgz` format files that you may have seen when downloading Linux programs from Internet sites. But `tar` and `cpio` can also create an archive directly on a tape cartridge or other back-up device without first creating a file on your hard disk.

TIP In order to use a tape drive or other back-up device, you must first install and configure that device using the information presented in previous chapters. For example, see Chapters 2, 3, and 4 regarding the installation of Linux and the use of kernel modules for adding device support.

The `tar` and `cpio` commands operate differently. With the `tar` command you must specify files to be included in a back-up archive on the command line. By contrast, `cpio` always looks in the STDIN channel for the filenames to include in an archive. The `tar` command writes data to a filename or device that you provide; the `cpio` command always writes data back to STDOUT. To compare these two methods of operation, consider the following two examples for creating a full backup of the `/home` directory. You can assume for this example that the device `/dev/tape` is configured as a tape drive. (Notice that you refer directly to a tape drive device; you do not mount it first.)

```
tar cf /dev/tape /home
```

This command uses the `c` option of `tar` to create a new archive. The `f` option (for *filename*) followed by the device name indicates the location where the archived data will be stored. The last parameter, `/home`, indicates which files will be archived. Because the parameter is a directory name, `tar` will include all files located within that directory. A `cpio` command equivalent to the above `tar` command would be:

```
find /home -print | cpio -o > /dev/tape
```

To use `cpio`, you must use the `find` command to generate a list of files (one filename per line) for `cpio` to refer to. The `find` command with the `-print` option generates this list. Those filenames are sent to `cpio` using a pipe symbol because `cpio` reads the filenames in from STDIN. The > redirection operator then sends the archived files to the device `/dev/tape`. The `-o` option on `cpio` indicates that the archive is being output—that is, that data is being written out. A simpler example of `cpio` could archive the contents of a single directory to a local file using the `ls` command to generate the list of files to archive:

```
ls | cpio -o /tmp/archive.cpio
```

TIP

You might have noticed that `tar` options do not normally include a preceding hyphen; those of the `cpio` command do.

The `v` option is normally added to both `tar` and `cpio` so that the output of the command is *verbose*, meaning that the command prints details of what it is doing to the screen. With that option added, the last example would look like this:

```
ls | cpio -ov /tmp/archive.cpio
```

Extracting files using `tar` or `cpio` is a similar operation, but using different options. If you had created an archive on a tape cartridge using `tar`, you could restore the contents of the tape into the current directory using this command (with the `x` option standing for *extract* and the `v` option included to see verbose messages about command progress):

```
tar xvf /dev/tape
```

The `cpio` command uses the `-i` option for input, again extracting the contents of the back-up media into the current directory. The `-d` option is also added here so that `cpio` will create subdirectories that existed in the data as required to re-create the original data organization. When using the `cpio` command with the `-i` option, `cpio` reads the STDIN channel to get the archived data; so the < redirection operator is used with the filename or archive device name.

```
cpio -idv < /dev/tape
```

These are very basic examples of `tar` and `cpio`. Each command supports dozens of options for features such as compressing files, preserving file attributes, controlling a tape device, setting timestamps on archived data, and many other things. You can review the manual and info pages for each command to learn more.

Both `tar` and `cpio` rely on other Linux commands to help you create an incremental or multilevel backup. The most useful of these is the `find` command. For example, the following `find` command will print a list of all files in the `/home` directory (and its subdirectories) that have been modified in the last day (note the `-mtime` parameter):

```
find /home -mtime 1 -print
```

By using the list of files generated by this command as the archive list for `cpio` or `tar`, you can easily create a level 2 backup, as described in the example earlier in the chapter, in which each level 2 backup contains all files modified since the last level 1 backup. (In this case, this command would be used on Tuesday; a different number of days would be used for each day of the week so that data changed since the beginning of the week was included in the backup.) The following two commands illustrate this:

```
find /home -mtime 1 -print | cpio -ov > /dev/tape
tar cf /dev/tape ` find /home -mtime 1 -print`
```

The options available with the `find` command make it a powerful companion to the `tar` and `cpio` commands. With `find` you can create a list of files owned by certain users, files modified or accessed within certain time limits, files with certain file permissions, or many other criteria.

Other Back-up Utilities

The `tar` and `cpio` commands can operate either with a tape drive or with back-up devices that rely on a standard `ext2`-style file system or standard mounting operation, such as a Jaz drive or a writeable CD drive. As mentioned earlier, tape drives are popular tools for back-ups, but they often require additional tools to manage tape indexes, tape rewinding and searching, and so forth. If you intend to use a tape drive, a freely available graphical utility worth reviewing is included with the KDE Desktop. The utility is called `kdat`, or the Tape Back-up Tool. It is included on the Utilities submenu under KDE menus in Red Hat Linux when using Gnome. Most other Linux distributions will include this program on the Utilities submenu of the KDE main menu.

The Tape Back-up Tool provides handy features like the following, all available from a graphical interface and menu structure (see Figure 14-5):

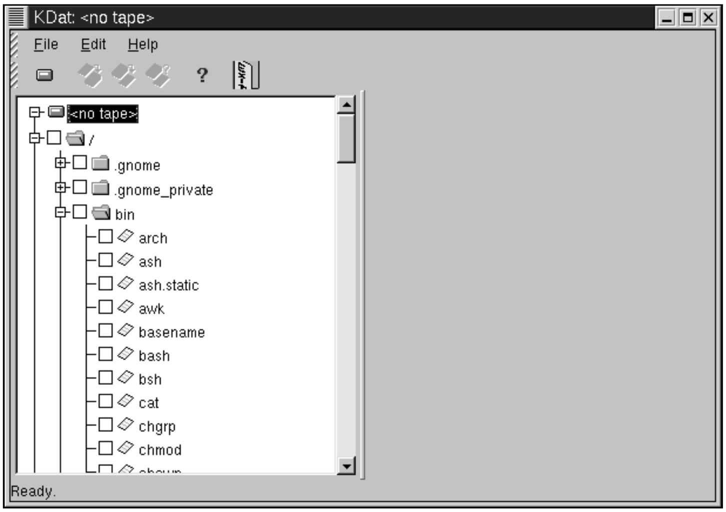

Figure 14-5 The Tape Back-up Tool in KDE

- Back up and restore files by dragging and dropping them between a list of the tape contents and a list of the hard disk contents
- Verify tape contents from the menu
- Manage mounting and unmounting of tape cartridges
- Create and maintain indexes of multiple tapes

- Set preferences from a graphical dialog box (see Figure 14-6)
- Format tapes

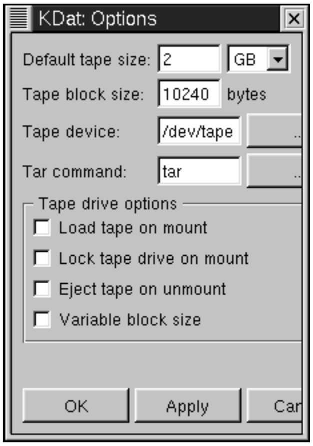

Figure 14-6 Setting preferences in the KDE Tape Back-up Tool

The Tape Back-up Tool is not intended to be compatible with all the high-end tape drives that you might consider using for your Linux servers, but it provides an easy-to-use method of tracking backups. It also makes it very simple to access data from a back-up tape.

Commercial Back-up Utilities

The complexities of maintaining large numbers of back-up media for large volumes of data led manufacturers long ago to create specialized software to help with the task. Fortunately, some of these tools have made their way to the Linux platform, and others appear to be forthcoming.

14

The best-known back-up utility with a strong following among Linux users is BRU, the back-up and restore utility, from Enhanced Software Technologies (see *www.bru.com*). The main

screen of BRU is shown in Figure 14-7, and the scheduling tool is shown in Figure 14-8. This product, which is included with some Linux distributions, provides features such as:

- Multiple levels of data verification
- Unattended operation with scheduled backups.
- Assistance in labeling large numbers of tapes, including backups that require multiple back-up media
- Support for numerous types of back-up devices

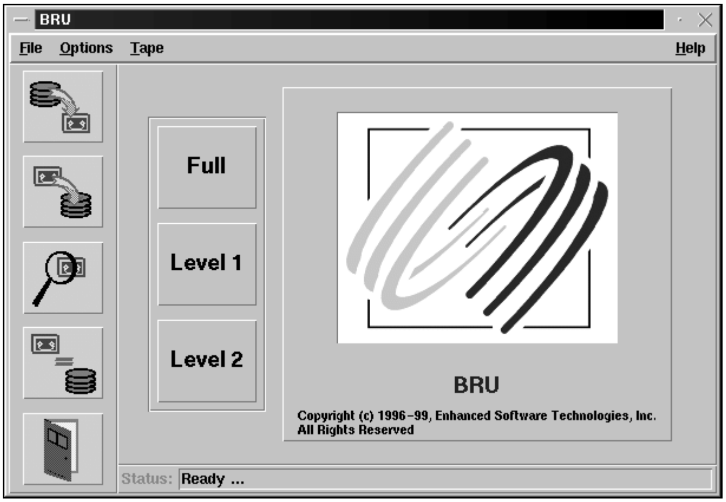

Figure 14-7 The main screen of BRU

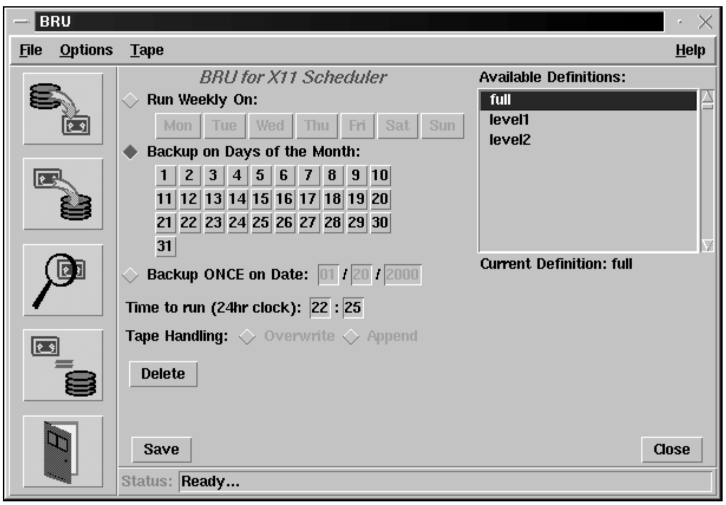

Figure 14-8 The scheduling tool in BRU

Another popular Linux back-up tool is Arkeia, from Knox Software (see *www.arkeia.com*). The Arkeia product is advertised as an enterprise network back-up solution and is considered a more full-featured tool than BRU. It is designed to control backup of multiple remote systems from a single location, saving or restoring data from anywhere on the network. Figure 14-9 shows a sample screen from the Arkeia program.

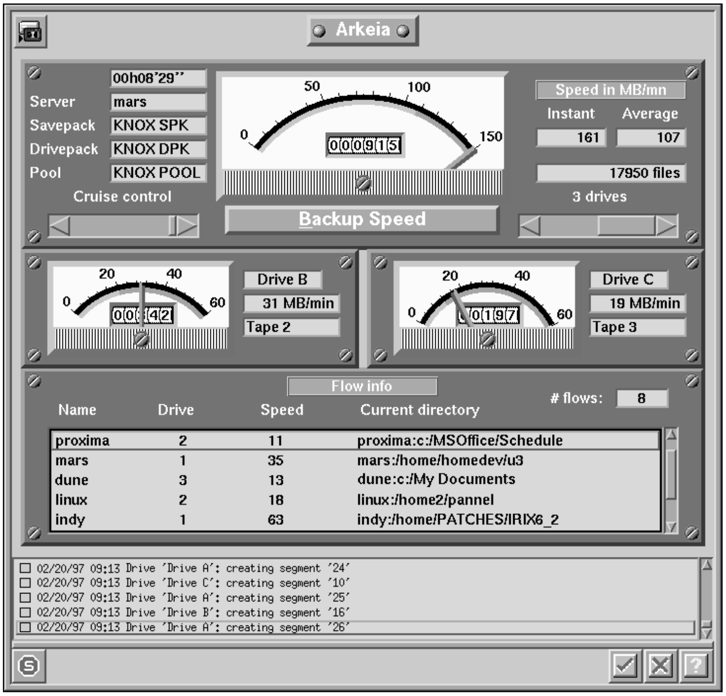

Figure 14-9 The Arkeia enterprise back-up program

Several additional back-up utilities are available for Linux or soon will be—some as part of high-end hardware platforms from companies such as MTI (*www.mti.com*) and Legato (*www.legato.com*).

CHAPTER SUMMARY

- Creating a back-up plan to safeguard an organization's data involves many considerations, such as the value of data, which devices and media formats are best suited to protecting that data, and how and when data should be backed up.
- Many types of back-up devices are available, with the most widely used being various tape cartridge formats. Optical devices have many advantages but lack the storage capacity of tape drives. Issues such as verifying data, compressing data, and restoring the root file system all must be considered when preparing the back-up plan.

- The `tar` and `cpio` utilities can be used to create simple backups, including incremental or multilevel backups. Numerous free and commercial graphical utilities are available to make back-up maintenance more convenient for a system administrator.

KEY TERMS

backup — A copy of data on a computer system.

back-up level — A definition of how much data is to be backed up in comparison with another back-up level. When performing a back-up operation at a given back-up level, all of the data that has changed since the last backup of the previous level is recorded.

back-up media — A device where data can be stored, such as a tape cartridge, writeable CD, or even a floppy disk.

back-up plan — A written document that outlines when, how, and, perhaps, why various files and file systems will be backed up, stored, and—when necessary—restored to prevent permanent data loss.

`cpio` — A Linux archiving program. The `cpio` command also reads archive files created by the `tar` command.

jukebox — A back-up device that holds multiple back-up media (such as multiple tape cartridges or writeable CDs) and that can switch between them without assistance from a system administrator.

legacy systems — Computer systems that an organization already owns. This term usually refers to systems that are no longer state of the art.

restore — To copy data from a back-up location (for example, a tape cartridge) onto the file system where that data is normally used, and from which it was unintentionally lost.

`tar` — A Linux archiving program.

REVIEW QUESTIONS

1. A back-up plan would normally not include the following:
   a. A list of tape drive prices
   b. Times when backups are performed
   c. The location of critical files on the system
   d. A recommended time to replace old tape cartridges with new ones
2. Explain how the speed with which files need to be restored affects a back-up plan.
3. It is important always to back up the operating system files as often as user data files. True or False?
4. Which of the following is part of measuring the value of data?
   a. The value of a project that cannot be done because data needed for the project was destroyed
   b. The cost of a complete set of back-up media to implement a three-level back-up plan

   c. The average data transfer rate of the chosen back-up device
   d. The average storage capacity of similar back-up devices
5. Name two parts of a Linux system that are likely to change daily.
6. Explain why a level 1 backup is called an incremental backup.
7. Using back-up levels has the advantage of:
   a. Reducing the time required to back up the entire file system
   b. Making it easier to recover a file that has not been changed in several weeks
   c. Allowing a system administrator to spend less time with backups but keep data backed up very frequently
   d. Causing all system backups to be available via a single file index
8. Using a standard three-level back-up plan with the time intervals described in the chapter text, a user would expect never to lose more than ______ worth of work.
   a. A week's
   b. A day's
   c. An hour's
   d. 20 MB
9. Explain in detail why a system administrator must use back-up media from three backups in order to completely restore a system that used three back-up levels.
10. Floppy disks are a useful back-up media in cases where:
    a. The cost of writeable CDs is prohibitive.
    b. Small amounts of critical data need to be backed up.
    c. Extreme durability is a key factor in the choice of media.
    d. A high data transfer rate is critical.
11. As a rule, tape cartridges can hold much more than optical media. True or False?
12. Describe two advantages of a back-up device with a jukebox feature.
13. Using a SCSI interface to connect a back-up device has the advantage of:
    a. Low cost
    b. Being proprietary (controlled by one company)
    c. High data transfer rates
    d. Limited availability
14. Name five factors to consider when selecting a back-up device and media type. Explain the circumstances in which each would be a controlling factor in the decision.
15. Name three tape cartridge formats and comment briefly on each.
16. You can expect a CD or other optical media to last about as long as high-quality microfilm. True or False?

17. The purpose of verifying your backups is to:
    a. Be certain that files are correctly recorded and can be restored
    b. Ascertain whether anyone has tampered with data contained in a backup
    c. Secure data from unauthorized use
    d. Compare data transfer rates among competing products
18. Explain how redundancy applies to compressed data.
19. The `tar` utility differs from the `cpio` utility in that:
    a. `cpio` always reads and writes to STDIN and STDOUT, while `tar` uses command-line parameters.
    b. `cpio` is a commercial utility, while `tar` is free software.
    c. `cpio` is widely used for Internet archive files, while `tar` is not.
    d. `cpio` is an older format that is not compatible with newer `tar` archives.
20. The ______________ utility is a commercial back-up utility from Knox Software.
    a. BRU
    b. kdat
    c. Arkeia
    d. mke2fs
21. Describe why the `find` command is often used with `tar` or `cpio` for incremental backups.
22. The ______________ option causes the `tar` command to extract files from an archive file or device.
    a. `a`
    b. `x`
    c. `c`
    d. `e`
23. Describe the special considerations that must be taken in order to restore the root file system of Linux after a hardware failure.
24. Name three removable media formats besides CD and tape cartridges.
25. In the long term, back-up media are likely to cost more than the back-up device used to access them. True or False?

Hands-on Projects

Project 14-1

In this activity you learn more about the Arkeia and BRU commercial back-up utilities. To complete this activity you should have a Web browser with access to the Internet.

1. Start your Web browser and go to **www.arkeia.com**.
2. Review the Supported Platforms page. What comments would you make about this company's support of Linux? What advantages do you foresee if you choose to use this software in a large organization that uses many types of computers?
3. Review the Product Features pages. (There are many pages with different categories of features.) Locate three features that you understand, based on what you have learned in this chapter. (You will not understand all of the features after reading this chapter.)
4. Change to the section of the Web site containing white papers (technical reports). Select one of the reports and read it online.
5. If you are interested in experimenting with this software, download a copy.
6. Go to the BRU Web site at **www.bru.com**.
7. Explore the Products and Support pages. What comments do you have about the differences in the two products?
8. If you are interested in experimenting with this software, download a copy using the **Download** link.

Project 14-2

In this project you explore the Web site of a major computer sales company to learn more about what back-up devices are available. To complete this activity you should have a Web browser with access to the Internet.

1. Start your Web browser and go to **www.warehouse.com**.
2. Verify that the PC Products page is displayed, and then choose the **Drives/Storage** link.
3. Under the Removable Storage heading on the left side of the browser window, choose the **Tape** link.
4. Review the "best-selling" devices pictured. Do you recognize the formats from those mentioned in the chapter? Do you recognize some of the brand names of the manufacturers?
5. Under PC Tape Drives on the left side of the browser window, choose **Advanced Intelligent Tape (AIT)**.
6. Review the brands, capacities, and prices presented in this category of products.
7. Use the browser's Back button to return to the previous page.
8. Under PC Tape Drives on the left side of the browser window, choose **Travan**.
9. Review the brands, capacities, and prices presented in this category of products. How do they compare with the AIT devices?

10. In the banner at the top of the screen, click on the **Supplies** link.
11. Under the Supplies heading on the left side of the browser window, choose **Magnetic Media**.
12. Under the Magnetic Media heading on the left side of the browser window, choose **QIC 1/4" Tape**. (This is the tape used by Travan-format drives.) Note the prices of the tapes.
13. Use the Back button on your browser and explore the prices of other tape formats.

Project 14-3

In this activity you use the `tar` command to create a simple data archive file and then extract the contents of that file into another directory. To complete this activity you should have a working Linux system with `root` access.

1. Log in to Linux as `root`.
2. If you logged in using a graphical login window, open a command-line window.
3. Enter **`cd /etc`** to change to the `/etc` directory.
4. Enter **`ls -l | less`** and review the filenames and file permissions that you have to the various configuration files in this directory.
5. Create a `tar`-format archive of the configuration files in the `/etc` directory using the command **`tar cf /tmp/testing.tar /etc`**. Because you are including the pathname to both the `testing.tar` archive file and the directory containing the information you want to archive, you could execute this command from any location on the system.
6. Enter **`tar cvf /tmp/testing2.tar /etc`**, which is a similar command, this time including the `v` option. After you execute this command, you see a list of all the files in the `/etc` directory appear on the screen as each is added to the archive file.
7. Change to your home directory by entering the **`cd`** command.
8. Use the **`ls`** command to examine the contents of your home directory. Make certain you do not have a file called `lilo.conf` in your home directory. (You shouldn't, but if you do from a previous exercise, rename it to something else to complete this project.)
9. Enter **`tar xvf /tmp/testing.tar etc/lilo.conf`** to use the `x` option of the `tar` command to extract a single file from the `tar` archive that you just created. The file is placed in your current directory. Notice that because of the `v` option the filename is printed to the screen as it is extracted.
10. Use the **`ls`** command to review the contents of your home directory. Do you see a file called `lilo.conf`? Look for an item named `etc`.
11. Enter **`cd etc`** (without a forward slash) to change to the `etc` subdirectory of your home directory. The `tar` command created the subdirectory in which the requested file was located, starting with your current directory when you issued the command to extract the file from the archive.
12. Use the **`ls`** command again to see the `lilo.conf` file in the `etc` subdirectory of your home directory.

CASE PROJECTS

1. You are working for General Linux Corporation, a relatively new company focused on Linux products and services. The company has about 100 employees and has just become a publicly traded corporation. As news of the public stock offering becomes more widely known, the workload on all employees is increasing, but everyone is pleased with the opportunities provided by this step. The company's Web site is also extra busy as potential customers and investors review information about the company and its products. As a system administrator, do you think the three-level back-up strategy outlined in the chapter text is sufficient to protect the data stored on company servers? What changes would you make to that plan?
2. Given that the company considers itself an "Internet company," with many large Linux servers and all employees working with Internet resources each day, which of the back-up utilities you have reviewed would you consider using? Why?
3. What factors influence your decision about which devices and media to use for your back-up plan at General Linux? What factors do you consider in valuing the data that you are protecting via your back-up plan?